HHS Public Access

Author manuscript

J Nutr. Author manuscript; available in PMC 2016 March 29.

Published in final edited form as:

J Nutr. 2013 June ; 143(6): 948S–956S. doi:10.3945/jn.112.172957.

Regression modeling plan for twenty-nine biochemical indicators of diet and nutrition measured in NHANES 2003–2006[1–,3]

Maya R. Sternberg, **Rosemary L. Schleicher**, and **Christine M. Pfeiffer***

National Center for Environmental Health, CDC, Atlanta, GA

Abstract

The collection of papers in this journal supplement provides insight into the association of various covariates with concentrations of biochemical indicators of diet and nutrition (biomarkers), beyond age, race and sex using linear regression. We studied 10 specific sociodemographic and lifestyle covariates in combination with 29 biomarkers from NHANES 2003–2006 for persons ≥20 y. The covariates were organized into 2 chunks, sociodemographic (age, sex, race-ethnicity, education, and income) and lifestyle (dietary supplement use, smoking, alcohol consumption, BMI, and physical activity) and fit in hierarchical fashion using each chunk or set of related variables to determine how covariates, jointly, are related to biomarker concentrations. In contrast to many regression modeling applications, all variables were retained in a full regression model regardless of statistical significance to preserve the interpretation of the statistical properties of *beta* coefficients, *P*-values and CI, and to keep the interpretation consistent across a set of biomarkers. The variables were pre-selected prior to data analysis and the data analysis plan was designed at the outset to minimize the reporting of false positive findings by limiting the amount of preliminary hypothesis testing. While we generally found that demographic differences seen in biomarkers were over- or under-estimated when ignoring other key covariates, the demographic differences generally remained statistically significant after adjusting for sociodemographic and lifestyle variables. These papers are intended to provide a foundation to researchers to help them generate hypotheses for future studies or data analyses and/or develop predictive regression models using the wealth of NHANES data.

INTRODUCTION

A vast amount of data is collected on each sampled person in the continuous NHANES survey, providing a unique opportunity to assess and describe the nutritional status of the US

[1]No specific sources of financial support. The findings and conclusions in this report are those of the authors and do not necessarily represent the official views or positions of the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry or the Department of Health and Human Services.

[3]Supplemental Tables 1–2 and Supplemental Text 1–2 are available from the “Online Supporting Material” link in the online posting of the article and from the same link in the online table of contents at http://jn.nutrition.org.

*To whom correspondence should be addressed: Christine M. Pfeiffer, Division of Laboratory Sciences, National Center for Environmental Health, Centers for Disease Control and Prevention, 4770 Buford Hwy, NE, Mail Stop F43, Atlanta, GA 30341, Phone: 770-488-7358, CPfeiffer@cdc.gov.

[2]Author disclosures: M.R. Sternberg, R.L. Schleicher, and C.M. Pfeiffer, no conflicts of interest.

population. However, NHANES cannot assess cause and effect. The variables collected in observational studies, such as NHANES, have not been experimentally manipulated and/or randomly assigned. In this setting any causal pathway becomes obfuscated and differences may provide little insight into the cause and effect. Observational studies, however, can still provide an approximate description of patterns in the data and form a basis to estimate associations and perform hypothesis testing after controlling simultaneously for many variables, though estimates may always be biased due to residual or unmeasured confounding.

Application of any statistical method first requires a well-formulated problem within the scope of the study design's ability to provide solutions. Adhering to the tenets of the scientific method should precede any statistical analysis. While the basic assumptions of the statistical method remain important, uncritical application and/or repeated application of a statistical modeling analysis without a well-formulated plan can simply capitalize on the random variation and lead to a model that has little utility for prediction, statistical estimation or testing, and rather leads to false positive findings (1, 2, 3).

One of the hallmarks of the scientific method is a "feedback loop" between theory and practice as we further refine our hypotheses after accumulating new facts (4). NHANES can be used to inform the feedback loop by providing a description of the nutritional status of the US population by various demographic, socioeconomic, health, and risk markers; and with its continuous design, the snapshots reflect changes in the nutritional status of the US population. This information can be used to develop modified hypotheses to further understand the reasons for observed differences and to identify important factors to consider when designing new experimental studies. The collection of papers in this journal supplement provides a systematic description of various biochemical indicators of diet and nutrition using the same sets of pre-defined covariates. These go beyond age, sex, and race-ethnicity, which have already been described in the CDC's *Second National Report on Biochemical Indicators of Diet and Nutrition in the US Population* (5). The main objective of this paper is to discuss the statistical strategy used to analyze the 10 selected sociodemographic and lifestyle correlates of nutritional biomarker concentrations belonging to different classes of nutrients and the reasons for limiting data-driven decisions commonly applied in many other NHANES analyses. A secondary objective was to summarize parameters from the model results presented in the accompanying papers to identify any general patterns.

SUBJECTS AND METHODS

Biomarkers

The dependent variables in the papers in this journal supplement include biomarkers of diet and nutrition measured in adults ≥20 y who provided a biological specimen during their examination at the mobile examination center in NHANES 2003–2006. Some biomarkers were only available for a subset of the full sample or for only 2 of the 4 survey years, i.e. 1 cycle (Table 1). It is known that a log transformation of data derived from biological assays can often be used to make the distribution of the data approximately normal. Previous analysis of NHANES biomarker data (5, 6) used simple graphical methods such as normal

probability plots and histograms to confirm the adequacy of a natural log transformation for the biomarkers presented. The natural log transformation was used for all the biomarkers considered in this study, with the exception of vitamin C, 25-hydroxyvitamin D, and body iron. The use of the logarithm in linear regression provides a straightforward interpretation of the response that is not in the units of the original response variable, but as a percent change. Thus, one can compute the percent change in response at 2 different values of a covariate, while holding all others constant (see Supplemental Text 1). In addition, with the aid of the Taylor series approximation, $\ln(x+1) \approx x$ for small x, the *beta* coefficient (multiplied by 100) from the multiple linear regression of a natural log transformed response can be approximately interpreted as the percent change in the response for a 1 unit change in the covariate (assuming the covariate has not been transformed), while holding all other variables constant (see Supplemental Text 2 for an example). Similarly, if both the response and the covariate have been transformed by the natural log, one can approximately interpret the *beta* coefficient as the percent change in the response for every 1 percent change in the covariate.

Composite variables are often used in public health messaging and the scientific literature. For fat soluble biomarkers, the following composite variables were created by summing a group of chemically related compounds: carotenes, xanthophylls, saturated, monounsaturated, polyunsaturated, and total fatty acids. These composite variables were only calculated for persons who had non-missing values across all corresponding biomarkers. Therefore, a small number of values were missing for these composite variables relative to the individual biomarkers (see Supplemental Table 1).

Covariates

Ten specific sociodemographic and lifestyle factors were selected as covariates based on the information available in NHANES and on evidence in the literature that these variables may be related to nutritional biomarkers. The sociodemographic variables included age, sex, race-ethnicity, education level, and family poverty income ratio (PIR[4]). For bivariate analyses, we categorized the sociodemographic variables as follows: age (20–39 y, 40–59 y, and ≥60 y); race-ethnicity (Mexican American [MA], non-Hispanic black [NHB], and non-Hispanic white [NHW]); education (<high school, high school, and >high school); PIR was calculated by dividing total family income by the poverty guidelines adjusted for family size at year of interview (7) and categorized as low (0–1.85), medium (>1.85–3.5), or high (>3.5), using the 1.85 cutoff that corresponds to income-eligible for the Special Supplemental Program for Women, Infants, and Children (8). The lifestyle factors included dietary supplement use, smoking, alcohol consumption, BMI, and physical activity level. Participants were categorized as “smokers” if their serum cotinine concentration was >10 μg/L (9). For descriptive tables, BMI (kg/m^2) was categorized using WHO guidelines for underweight (<18.5), normal (18.5–<25), overweight (25–<30) and obese (≥30) (10). Assessment of supplement use was based on whether the participant indicated any use of dietary supplements in last 30 d.

[4]Abbreviations used: MA, Mexican American; MET, metabolic equivalent task; NHB, non-Hispanic black; NHW, non-Hispanic white; PIR, poverty income ratio

The physical activity variable was constructed using files that provide detailed information about specific leisure time physical activities (11). Participants were asked to recall the frequency, duration, type and intensity of each leisure time physical activity for which they engaged for more than 10 min during the past 30 d. To construct a physical activity variable that accounts for energy expenditure, the metabolic equivalent task (MET) score for each leisure time physical activity was multiplied by the frequency and duration (min). This variable was then summed for each participant, divided by 30 and then multiplied by 7 to obtain total MET-min/wk. For the descriptive tables, this variable was categorized into 4 groups: no reported leisure time physical activity, 0–<500, 500–<1000, and ≥1000 MET-min/wk (12).

Average daily alcohol consumption was derived from the alcohol use questionnaire as: [(quantity × frequency) / 365.25]. Respondents were asked about their alcohol use where a drink was defined as a 12 oz. beer, a 5 oz. glass of wine, or 1.5 oz of liquor. This is equivalent to a “standard” drink in the United States, which contains 0.6 US fluid oz (18 mL) of alcohol and corresponds to 14.2 g of ethanol. Persons who reported having less than 12 drinks of any type of alcoholic beverage in the past year (or lifetime) were considered nondrinkers. For descriptive purposes, alcohol consumption was categorized in the following groups: no drinks, <1, 1–<2, and ≥2 drinks/d.

In a few cases additional variables were added to the full model (sociodemographic and lifestyle factors) to provide important adjustments that might be expected for certain biomarkers. For the urine biomarkers, urine creatinine concentration was included as a covariate to adjust for the dilution of the spot urine. For fat-soluble nutrients, total cholesterol and prescription use of lipid-altering drugs was included because some fat-soluble nutrients are transported in the plasma by lipids and to adjust for drug-related changes in fat absorption and/or lipid metabolism. For 25-hydroxyvitamin D, season and latitude were included as proxies to adjust for sun exposure, which has been shown to have an impact on vitamin D status.

The mathematical form of the continuous covariates (age, PIR, BMI, physical activity, and alcohol consumption) was assumed to be linear in the regression model. A log transformation for BMI, alcohol consumption, and physical activity was applied to these covariates; although not a necessary assumption, linear regression is more robust when the independent variables have an approximately normal distribution (13). To deal with 0 in alcohol consumption and physical activity data, a ln(x+1) transformation was applied (Table 2).

Statistical analyses that apply to accompanying papers in this supplement

The analysis plan entailed computing Spearman correlations to describe bivariate associations between each biomarker and selected continuous variables. Bivariate associations between each biomarker and categorical variables were described with geometric means (or arithmetic means where appropriate) and 95% CI across the categories. The means were compared across the categories on the basis of Wald F tests (tests whether at least 1 of the means across the categories is significantly different from the others).

Geometric or arithmetic means were not presented if the minimum sample size of 42 was not reached (assumed average design effect of 1.4 multiplied by 30) (14).

At the outset, we identified 10 different covariates for 29 different responses (biomarkers) and decided to keep all variables in a full regression model regardless of statistical significance. Covariates were arranged into 2 sets or "chunks" of sociodemographic factors and lifestyle factors. We tested these covariates in a hierarchical, chunk-wise fashion such that each chunk or set of related variables was tested simultaneously (15, 16). The influence of each chunk was assessed by a Satterthwaite adjusted F chunk test, which tests whether at least 1 of the model coefficients for the set of variables in the chunk is significantly different from 0. Wald F test statistics were used to test whether any single coefficient was significantly different from 0, if the overall chunk test was statistically significant. In addition to simple linear regression (model 1), multiple regression models were considered for each biomarker: a multiple linear regression model with the sociodemographic variables (model 2), and a multiple linear regression model with both the sociodemographic and lifestyle variables in the regression (model 3). For urine biomarkers, urine creatinine concentration was added to model 3 (model 4). For fat-soluble nutrients, lipid-related factors were added to model 3 (model 4) or sun exposure factors were added to model 3 for 25OHD only (model 5). Simple linear regression was used to provide an estimate of the unadjusted *beta* coefficient and a sample coefficient of determination (R^2). Assuming the model is not misspecified, confounding can be assessed by the change in estimate between the adjusted and unadjusted *beta* coefficients from these models (16). We assessed confounding in specific variables by noting the change in the *beta* coefficients from the simple linear regression (model 1) to a larger multiple linear regression model (models 2–5). The magnitude of change that constitutes confounding can vary by subject matter. In some situations, a relatively small change might be meaningful; whereas in other instances, a larger change might not be clinically meaningful. A rule of thumb is that any change in the *beta* coefficient greater than 10–20% may be considered confounding (16).

A factor that limited the number of *a priori* variables to consider was related to the available degrees of freedom (*df*). While thousands of people are sampled in any cycle of NHANES, the effective *df* available is based on the number of primary sampling units minus the number of strata (17). In NHANES this amounts to approximately 15 *df* per cycle. Many preliminary analytic decisions were made to ensure we had sufficient *df* to include all pre-selected variables simultaneously in a full model for each of the 29 analytes. Because several analytes only had 1 cycle of data released, we limited ourselves to a maximum of 15 *df*. In addition, we decided at the outset to exclude the consideration of higher-order interactions. Unless an interaction is driven by a known biological phenomenon, a statistically significant interaction will be difficult to interpret and in a descriptive analysis, such as this one, most likely represents general lack of model fit (15).

Statistical analyses were carried out using SAS for Windows software version 9.2 (SAS Institute, Cary, NC) and SAS-callable SUDAAN (SUDAAN Release 10.0, 2008 RTI, Research Triangle Park, NC) to account for the unequal probability of inclusion, stratification, and clustering. SUDAAN offers Taylor series linearization to account for the effect of stratification and clustering on the variance estimates. The weights used depended

on whether the specimens tested constituted a full or a subsample of all the eligible participants examined at the MEC and how many survey periods were combined to produce the estimate (Table 1). *P*-values < 0.05 were considered statistically significant. Because SUDAAN v10.0 does not have a correlation procedure, Spearman's correlations were computed as the slope of the regression of the standardized ranks for both variables. *P*-values for the Spearman correlation were computed as the maximum *P*-value of the slope coefficient of x on y and y on x.

Statistical analyses specific to this paper

We used a row-labeled plot to illustrate the increase in R^2 from model 2 (includes sociodemographic variables) to model 3 (includes both sociodemographic and lifestyle variables) for all 29 biomarkers. The arrows are sorted in ascending order based on the R^2 from model 2. Models 2 and 3 are nested and so the model with more covariates (model 3) will always have a larger R^2 than the smaller model (model 2).

To illustrate how controlling for more covariates selectively affects various biomarkers, we plotted the change in the *beta* coefficient (multiplied by 100) between model 1 and model 3 for age (for every 1 y increase), sex (females vs. males), and race-ethnicity (MA vs. NHW and NHB vs. NHW) for 20 of the 29 biomarkers using row-labeled plots. The arrow points in the direction of the change of the value of the *beta* coefficient from model 1 to model 3. Of the 29 biomarkers included in the analysis, 20 shared some critical properties that made them suitable for this comparison: (1) they are based on a natural log transformation facilitating interpretation and (2) they generalize to the adult population ≥20 y with no further restrictions. A *beta* coefficient of 0, suggests that a change in that covariate produces no change in the response. To provide information about broad patterns across the biomarkers for each of the covariates in models 3–4, a summary of the estimated adjusted percent changes is presented using the *beta* coefficients.

RESULTS

Among adults ≥20 y in the non-institutionalized, civilian US population in 2003–2006, 23% were ≥60 y, 52% were female, 72% were non-Hispanic white, 56% had more than a high school education, 43% were considered high income based on PIR, 29% had evidence of current smoking, 29% reported not having any alcohol consumption during the past year or ever, 54% reported taking dietary supplements, 33% were considered obese, and 32% reported no leisure time physical activities during the past 30 d that lasted more than 10 min (Table 3). As different biomarkers were analyzed in different NHANES survey periods and/or subsamples, we verified that this descriptive information was not qualitatively impacted by the set of NHANES weights used (see Supplemental Table 2).

It is interesting to note some patterns across the analytes when comparing the change in R^2 from model 2 to model 3 (Fig. 1). Urine phytoestrogens showed the smallest increase in R^2 after adding the lifestyle chunk; whereas both acrylamide and glycidamide showed the largest increases. Plasma homocysteine stood out as the analyte for which the sociodemographic variables explained most of the variability, yet with little added value from the lifestyle factors.

To illustrate how controlling for more covariates affects *beta* coefficients for demographic variables, we looked at the changes in *beta* coefficients for twenty analytes that use a natural log transformation and allow simplified interpretation of values (Fig. 2). Consider serum folate and the *beta* coefficient for sex with males as the reference. The *beta* coefficient from model 1 is 0.129 (95% CI: 0.104 – 0.154), while the *beta* coefficient from model 3 is 0.057 (95% CI: 0.031 – 0.083). Using the approximate interpretation, this suggests that females have approximately 12.9% (0.120 × 100) higher serum folate concentrations than males before controlling for covariates and 5.7% (0.057 × 100) after controlling for age, race-ethnicity, PIR, education level, smoking, alcohol consumption, BMI, physical activity, and supplement use. The change in the *beta* coefficient for sex and serum folate from model 1 to model 3 reveals a 56% change in the *beta* coefficient, which may imply that at least 1 of the variables in model 3 or a combination of them may have confounded the unadjusted estimate of model 1. On the other hand, some of the biomarkers reveal a qualitative change in the interpretation of the *beta* coefficient when the value changes from negative to positive or vice versa. Consider sex and acrylamide as an example. The *beta* coefficient for sex (with males as the reference) changes from −0.112 (95% CI: −0.149 – −0.0743) in model 1 to 0.0356 (95% CI: −0.0228 – 0.0940) in model 3. In other words, prior to any adjustment, acrylamide was approximately 11.2% lower in females compared to males. After adjusting, acrylamide was approximately 3.56% higher in females compared to males; in addition, the variable (sex) is no longer statistically significant after controlling for the remaining sociodemographic and lifestyle variables. On the other hand, for gylcidamide the unadjusted (model 1) and the sociodemographic adjusted (model 2) *beta* coefficient for sex are not significant, but once lifestyle factors are controlled for it reveals a statistically significant difference between females and males, such that females have approximately 8.7% higher levels of glycidamide than males. The reason for the sex gap cannot be explained by differences among the remaining variables, like smoking status, and may suggest other variables that have not been controlled for in the model such as genetics and other sex-specific effects that modify acrylamide metabolism.

To provide information about broad patterns across the biomarkers for each of the covariates in the full regression model, a summary of the estimated adjusted changes is presented (Table 4). While there are a few exceptions, the statistically significant associations between the biomarkers and age, sex, or race-ethnicity remained after adjusting for all the pre-selected covariates. In case of phytoestrogens, non-significant associations between age, sex or race-ethnicity became statistically significant or vice versa after adjusting for the pre-selected variables. Among the sociodemographic factors, education level had the fewest statistically significant associations among the biomarkers (5/29); age on the other hand was statistically significant for 24 and race-ethnicity for 22 (NHB vs. NHW) and 18 (MA vs. NHW) of the 29 biomarkers. Among these significant associations, age was most often positively associated, while race (NHB vs. NHW) was generally negatively associated with the biomarkers (Fig. 2, Table 4). The exceptions for age included pyridoxal-5'-phosphate, carotenes, 25-hydroxyvitamin D, and acrylamide hemoglobin adduct; the exceptions for race-ethnicity (NHB vs. NHW) included total cobalamin, vitamin C, carotenes, xanthophylls, soluble transferrin receptor, and glycidamide hemoglobin adduct. Both levels of comparison for race-ethnicity, NHB vs. NHW and MA vs. NHW, were simultaneously

statistically significant for 15 biomarkers; for 8 of the biomarkers NHW had the highest concentrations, for 3 biomarkers (total cobalamin, vitamin C and xanthophyll) NHW had the lowest concentrations. The remaining 4 biomarkers had associations for each level in opposite directions. Among the lifestyle factors, physical activity had the fewest statistically significant associations among the biomarkers (12 of the 29); whereas, BMI was statistically significant for 20 and smoking status for 19 of the 29 biomarkers. Out of the 15 biomarkers for which both smoking and BMI were statistically significant, the association was in the same direction for 12 (exceptions: RBC folate, soluble transferrin receptor, and acrylamide hemoglobin adduct). Similarly, out of the 10 times both smoking and alcohol consumption were statistically significant, the direction of the association was in the same direction for 8 (exceptions: pyridoxal-5'-phosphate and glycidamide hemoglobin adduct). The magnitude of the estimated change varied among the 16 statistically significant changes for supplement use; however, all the associations suggested increases in biomarker concentrations among supplement users, except for methylmalonic acid, total homocysteine, and acrylamide hemoglobin adduct. Supplement use was not significantly associated with any of the phytoestrogens and most fatty acids.

DISCUSSION

In developing a regression plan to assess the joint impact of 10 specific sociodemographic and lifestyle covariates for each of 29 biomarkers from NHANES 2003–2006 we tried to avoid some statistical practices that have been shown to capitalize on random variation such as repeated significance testing, data driven selection of optimal cut points for quantitative variables, automatic model selection approaches, and using the same data more than once to develop a regression model. Derksen and Keselman (18) showed through simulation, that the final model selected from stepwise selection included less than half of the actual number of real or true predictors; in addition, between 20–75% of the findings represented noise in the final model. Freedman (19) demonstrated through simulation and asymptotic theory that screening variables in a full regression model based on statistical significance followed by eliminating those that are not statistically significant could lead to models with high R^2 despite the fact that none of the covariates are truly related to the response. While the R^2 may not be the primary statistic of interest, the implication of an inflated R^2 is a small mean square error leading to inflated test-statistics for the *beta* coefficients and hence *P*-values that are more likely to reject the null hypothesis of no association. Another practice that we chose to avoid in this data analysis was to use the response data to determine the form of a continuous covariate in a regression model. This practice has been associated with an inflation of type I error when preliminary tests for non-linearity are used (20) and can lead to exaggeration of effect sizes, and smaller *P*-values (21). Recognizing there is a penalty associated with data mining (22), we decided in advance how to spend the available *df* and to avoid further model refinement. There is always a tension between bias and simplicity when developing a model. But one of the primary problems with developing a regression model using an iterative approach is that by the act of repeated hypothesis testing and data driven model decisions, one fails to preserve the statistical properties and interpretation of *beta* coefficients and standard errors that underlie the frequentist methods so often used in empirical research (3, 15, 18–23). Additionally, estimates of the standard error from a given

model account for the sampling variability assuming the model is true. They do not account for the uncertainty associated with not knowing which model is true. Thus, by capitalizing on the random variation researchers may develop models that show more agreement with the sample than with data on the entire population, or any other sample from that population (24).

This problem in model building has long been recognized and stems from the fact that we use the same data twice. Chatfield (25) writes “It is indeed strange that we often admit model uncertainty by searching for the best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true.” The benefits of limiting the number of hypothesis tests by keeping *a priori* selected variables in a regression model despite statistical significance include: providing findings that are more likely to be reproducible, preserving the interpretation of the full model results, and keeping the interpretation consistent across a set of biomarkers without over-interpreting the results for any single biomarker.

A key factor in determining the accuracy of the estimate, both its bias and its variance, is how well confounding has been controlled for in the model (15). Because our model does not control for all variables that may be important and/or confound the observed relationship, one must recognize this as a limitation of our analyses and inferences. A more parsimonious model could be derived for each biomarker by eliminating the variables that were not statistically significant from each chunk. However, we were more concerned with minimizing the penalty of data mining, preserving interpretation of the model results, and limiting the number of false positive associations, so we chose to report the results of the full model.

While one of the primary advantages of limiting data driven decisions during the model building process is the preservation of the statistical properties of *P*-values and confidence levels, it does not solve the problems associated with model misspecification or error-in-variable problems i.e., covariate measurement errors. There are many opportunities for model misspecification in our regression models. For example, the assumption of linearity between the continuous covariates and the biomarker may not be accurate; variables included in the model may be measured with error or may be suboptimal in other ways. For example, the dietary supplement use covariate does not specify which specific kinds of supplements are being used nor does it differentiate between persons who used supplements infrequently from persons who used them daily. Many important correlates of the individual biomarkers considered in our analyses were not included in our modeling approach, such as specific dietary intake variables. In addition, the type of strategy we employed is not immune from overfitting. The model could be specified, *a priori*, as too complex or prone to numerical instability by including too many covariates for the available *df* and/or including highly correlated covariates.

When analyzing data from observational studies, there are many legitimate reasons to explore and examine data in advance of creating a regression model; such as, error checking to confirm the integrity of the data, assessing the size of the sample and types of variables, checking for possible influential observations or the degree of missing data, or using

graphical methods to assess basic statistical assumptions (2). A data analysis plan could be developed having full access to all the variables except the outcome of interest with limited penalty (26). In addition, in some situations there may be ways to mitigate the impact of data driven decisions either by using methods that adjust for multiple comparisons, by adjusting the level of significance to account for preliminary testing, and/or by developing models that account for model uncertainty. However, developing regression models with complex survey data is challenging for many reasons, one of which is that many of the solutions to address some of the problems of model uncertainty have not been well-researched or implemented in commercially available statistical software in the context of complex survey data analysis. For example, a Bayesian solution to model uncertainty has been described by averaging across all the competing models and attaching weights of plausibility to each of the models and thereby incorporating a notion of model uncertainty (25), rather than identifying a single 'best' model. Other solutions propose cross-validation, shrinkage, penalized maximum likelihood estimation and resampling methods (15, 22). Regression modeling in complex surveys can be approached from a super-population inference point of view (17), which can more easily be adapted to some of the proposed solutions to assess model uncertainty. However, the publicly released NHANES data sets do not necessarily provide enough information to account for the sampling design, nor is it obvious how to incorporate all the design aspects into a model including fixed and/or random effects. In addition, the problem of causality is difficult in observational studies without very careful consideration of the causal pathway between exposure and outcome. In order to reasonably establish cause and effect, statistical methods like propensity scores that try to approximate the design of a randomized clinical trial (26, 27) are better suited than traditional regression methods.

In an effort to compare the effects of a fixed set of covariates across all biomarkers presented in the *Second National Report on Biochemical Indicators of Diet and Nutrition in the US Population* (5), we chose to forfeit insight into the association between variables unique to each individual biomarker and rather chose an approach that was consistent across a set of biomarkers and limited the amount of data mining. The approach taken in these papers provide a natural way for other researchers to build upon our results by selecting additional variables for a specific biomarker and adding additional chunks, such as a health status chunk and/or a dietary intake chunk or the researcher may choose to use a different variable elimination strategy. In summary, while we do not claim that any of our final models are 'correct', we have adhered to the scientific method and "rules of behavior" (28), focusing on the "feedback loop" (4) between theory and practice, to make decisions before fitting any regression models. Hence, the purpose of this set of papers is to provide an inductive foundation for researchers to build on these NHANES analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge contributions from the following individuals: Bridgette Haynes and Yi Pan. C.M.P, M.R.S., and R.L.S designed the overall research project; M.R.S., C.M.P, R.L.S, and M.E.R conducted most of the

research; M.R.S. analyzed most of the data and wrote the initial draft, which was modified after feedback from all coauthors, and had primary responsibility for content. All authors read and approved the final manuscript.

Literature Cited

1. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2:696–701.
2. Chatfield C. Avoiding statistical pitfalls. Stat Sci. 1991; 6:240–252.
3. Young SS, Karr A. Deming, data and observational studies: A process out of control and needing fixing. Significance. 2011; 8:116–120.
4. Box GEP. Science and statistics. J Am Stat Assoc. 1976; 71:791–799.
5. U.S. Centers for Disease Control and Prevention. Second National Report on Biochemical Indicators of Diet and Nutrition in the U.S. Population 2012. Atlanta, GA: National Center for Environmental Health; 2012 Apr. Available from: http://www.cdc.gov/nutritionreport [cited 2013 Feb 1]
6. U.S. Centers for Disease Control and Prevention. National Report on Biochemical Indicators of Diet and Nutrition in the U.S. Population 1999–2002. Atlanta, GA: National Center for Environmental Health; 2008 Jul. Available from: http://www.cdc.gov/nutritionreport/99-02 [cited 2013 Feb 1]
7. U.S. Department of Health & Human Services. Poverty Guidelines, Research, and Measurement. Washington, DC: U.S: Department of Health & Human Services; 2011 Jan.
8. USDA Food and Nutrition Service. [cited 2013 Feb 2] How to Apply: WIC Eligibility Requirements. Available from: http://www.fns.usda.gov/wic/howtoapply/eligibilityrequirements.htm
9. Bernert JT Jr, Turner WE, Pirkle JL, Sosnoff CS, Akins JR, Waldrep MK, Ann Q, Covey TR, Whitfield WE, et al. Development and validation of a sensitive method for determination of serum cotinine in smokers and nonsmokers by liquid chromatography/atmospheric pressure ionization tandem mass spectrometry. Clin Chem. 1997; 43:2281–2291. [PubMed: 9439445]
10. WHO. WHO Technical Report Series 894. Geneva: World Health Organization; 2000. Obesity: preventing and managing the global epidemic. Report of a WHO Consultation.
11. National Center for Health Statistics, Centers for Disease Control and Prevention. NHANES 2003–2004 data documentation: physical activity individual activities file. Hyattsville, MD: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2005. Available from: http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/PAQIAF_C.htm [cited 2013 Feb 1]
12. Wang CY, Haskell WL, Farrell SW, LaMonte MJ, Blair SN, Curtin LR, Hughes JP, Burt VL. Cardiorespiratory fitness levels among US adults 20–49 years of age: findings from 1999–2004 National Health and Nutrition Examination Survey. Am J Epidemiol. 2010; 171:426–435. [PubMed: 20080809]
13. Box GEP, Wastson GS. Robustness to non-normality of regression tests. Biometrika. 1962; 49:93–106.
14. National Center for Health Statistics, Centers for Disease Control and Prevention. Analytic guidelines and reporting guidelines: NHANES III (1999–1994). Hyattsville, MD: Centers for Disease Control and Prevention, US Department of Health and Human Services; 1996. Available from: http://www.cdc.gov/nchs/data/nhanes/nhanes3/nh3gui.pdf [cited 2013 Feb 1]
15. Harrell, FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer Verlag; 2001.
16. Kleinbaum, DG.; Kupper, LL.; Nizam, A.; Muller, KE. Applied regression analysis and other multivariable methods. 4th. Belmont: Duxbury; 2008.
17. Korn, EL.; Graubard, BI. Analysis of Health Surveys. New York: Wiley; 1999.
18. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol. 1992; 45:265–282.
19. Freedman DA. A note on screening regression equations. Am Stat. 1983; 37:152–155.
20. Grambsch PM, O'Brien PC. The effects of transformations and preliminary tests for non-linearity in regression. Stat Med. 1990; 10:697–709. [PubMed: 2068422]
21. Altman DG, Lausen B, Sauerbrei W, Schumacher M. The dangers of using 'optimal' cutpoints in analysis of quantitative exposures. J Natl Cancer Inst. 1994; 86:829–835. [PubMed: 8182763]

22. Faraway JJ. On the cost of data analysis. J Comput Graph Stat. 1992; 1:213–229.
23. Hurvich CM, Tsai C-L. The impact of model selection on inference in linear regression. Am Stat. 1990; 44:214–217.
24. Miller, AJ. Subset Selection in Regression. London: Chapman and Hall; 1990.
25. Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc Ser A Stat Soc. 1995; 158:419–466.
26. Rubin DB. The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Stat Med. 2007; 26:20–36. [PubMed: 17072897]
27. Rosenbaum PR, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.
28. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A Math Phys Sci. 1933; 231:289–337.

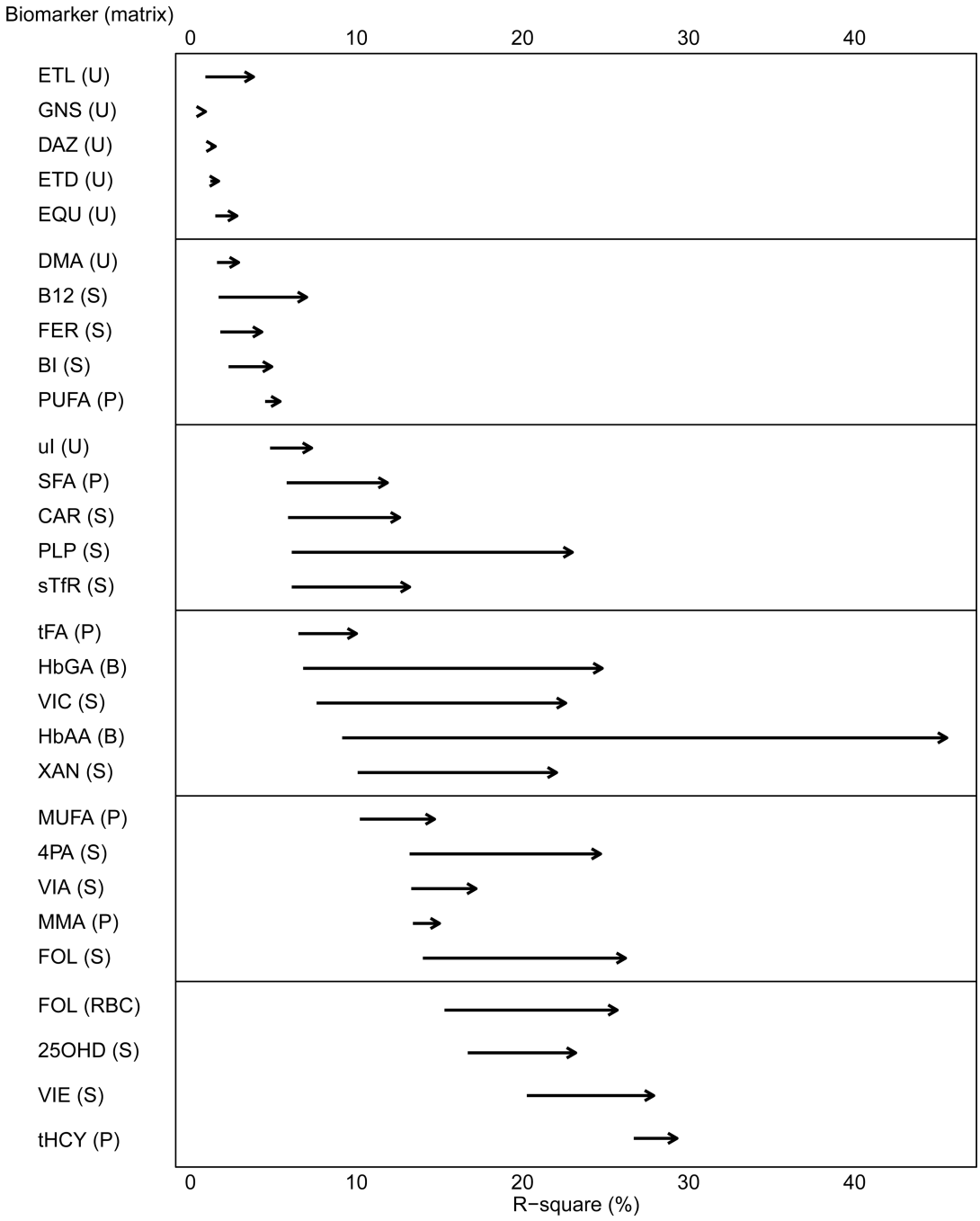

Figure 1.

Increase in R^2 from the multiple linear regression model with the sociodemographic variables (model 2) to the multiple linear regression model with the sociodemographic and lifestyle variables (model 3)

Sorted in ascending order based on model 2 R^2 (%); arrows point in the direction of the increase from model 2 to model 3 R^2; to simplify visual appearance, horizontal lines have been added.

25OHD, 25-hydroxyvitamin D ; 4PA, 4-pyridoxic acid; B, whole blood; B-12, total cobalamin; BI, body iron; CAR, carotenes [sum of *alpha*-carotene, *beta*-carotene and *cis*- and *trans*-lycopene], DAZ, daidzein ; DMA, O-desmethylangolensin; EQU, equol; ETD, enterodiol; ETL, enterolactone; FER, ferritin; FOL, folate; GEN, genistein; HbAA, acrylamide hemoglobin adduct; HbGA, glycidamide hemoglobin adduct; MMA, methylmalonic acid; MUFA, sum of 6 monounsaturated fatty acids; P, plasma; PLP, pyridoxal-5'-phosphate; PUFA, sum of 11 polyunsaturated fatty acids; S, serum; SFA, sum of 6 saturated fatty acids; sTfR, soluble transferrin receptor; tFA, total fatty acids [sum of 24 fatty acids]; tHcy, total homocysteine; U, urine; uI, urine iodine; VIA, retinol; VIC, ascorbic acid; VIE, *alpha*-tocopherol; XAN, xanthophylls [sum of lutein, zeaxanthin and *beta*-cryptoxanthin].

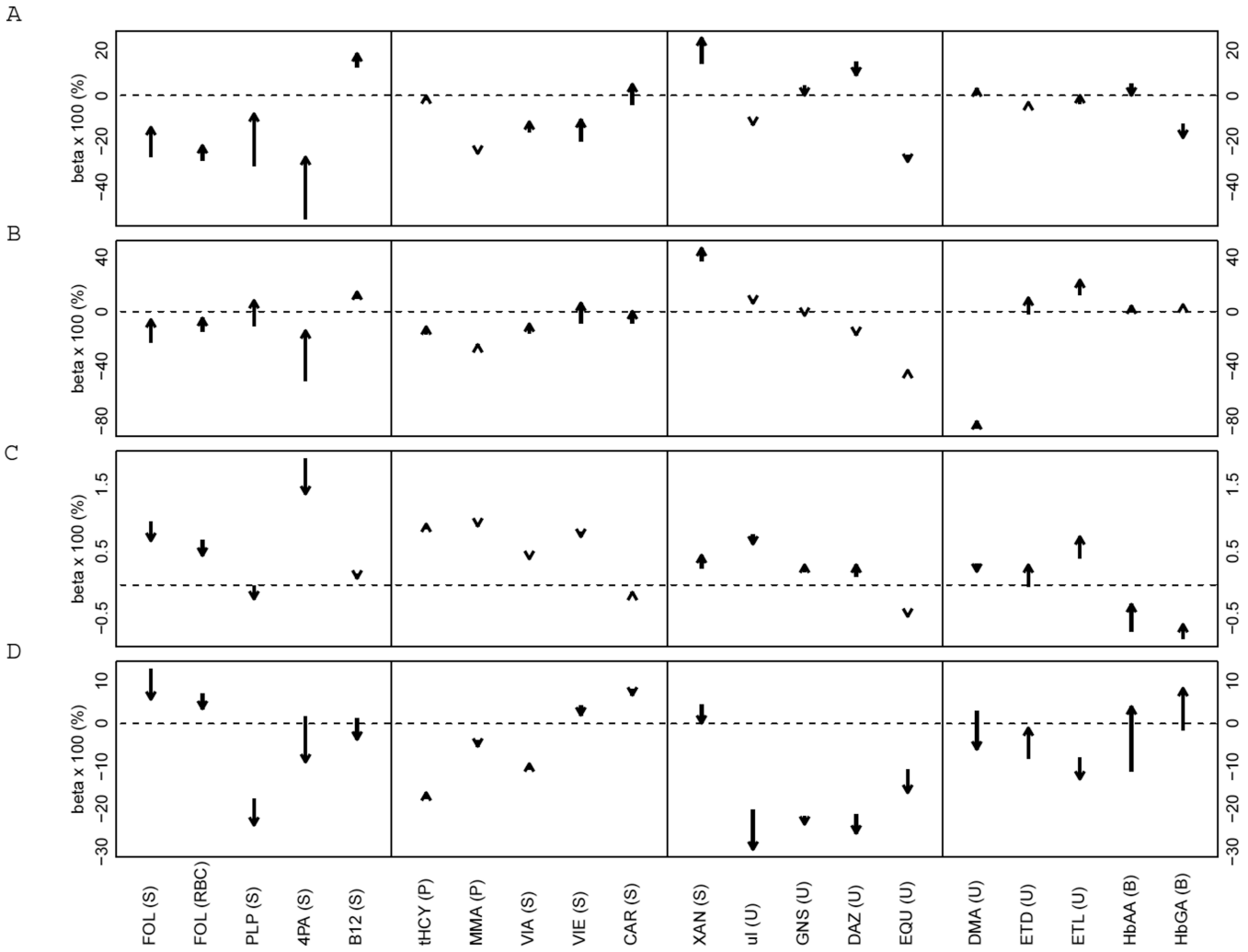

Figure 2.

Relative change in *beta* coefficient (multiplied by 100) for sex, age, and race-ethnicity from the simple linear regression (model 1) to the multiple linear regression model with the sociodemographic and lifestyle variables (model 3)

A: non-Hispanic black vs. non-Hispanic white; B: Mexican American vs. non-Hispanic white; C: 1 y increase in age; D: females vs. males.

In each panel, *beta* × 100 (%) can be interpreted as the approximate percent change in the biomarker for a change in the respective covariate while holding any other variables in the model constant.

Sorted by class of biomarker (water-soluble, fat-soluble, phytoestrogens, iodine, hemoglobin adducts of acrylamide); arrows point in the direction of the change of the *beta* coefficient from model 1 to model 3; reference line at zero; to simplify visual appearance, horizontal lines have been added.

4PA, 4-pyridoxic acid; B, whole blood; B-12, total cobalamin; CAR, carotenes [sum of *alpha*-carotene, *beta*-carotene and *cis*- and *trans*-lycopene], DAZ, daidzein ; DMA, O-desmethylangolensin; EQU, equol; ETD, enterodiol; ETL, enterolactone; FOL, folate; GEN, genistein; HbAA, acrylamide hemoglobin adduct; HbGA, glycidamide hemoglobin adduct;

MMA, methylmalonic acid; P, plasma; PLP, pyridoxal-5'-phosphate; S, serum; tHcy, total homocysteine; U, urine; uI, urine iodine; VIA, retinol; VIE, *alpha*-tocopherol; XAN, xanthophylls [sum of lutein, zeaxanthin and *beta*-cryptoxanthin].

Table 1

Biomarkers of diet and nutrition assessed in the adult US population ≥20 y during all of part of NHANES 2003–2006

Class	Biomarkers[1] (matrix[2])	Survey cycle	Population studied	Sample
Water-soluble	FOL (S), FOL (RBC), B-12 (S), tHcy (P), VIC (S)	2003–2006	≥20 y	Full
	MMA (P)	2003–2004	≥20 y	Full
	PLP (S), 4PA (S)	2005–2006	≥20 y	Full
Fat-soluble	VIA (S), VIE (S), CAR (S), XAN (S)	2005–2006	≥20 y	Full
	25OHD (S)	2003–2006	≥20 y	Full
	SFA (P), MUFA (P), PUFA (P), tFA (P)	2003–2004	≥20 y	Fasted subsample
Trace elements	FER (S), sTfR (S), BI (S)	2003–2006	Women 20–49 y	Full
	uI (U)	2003–2006	≥20 y	1/3 Subsample
Phytoestrogens	GEN (U), DAZ (U), EQU (U), DMA (U), ETD (U), ETL (U)	2003–2006	≥20 y	1/3 Subsample
Acrylamide	HbAA (B), HbGA (B)	2003–2004	≥20 y	Full

[1] 25OHD, 25-hydroxyvitamin D; 4PA, 4-pyridoxic acid; B-12, total cobalamin; BI, body iron; CAR, carotenes [sum of *alpha*-carotene, *beta*-carotene and *cis*- and *trans*-lycopene], DAZ, daidzein; DMA, O-desmethylangolensin; EQU, equol; ETD, enterodiol; ETL, enterolactone; FER, ferritin; FOL, folate; GEN, genistein; HbAA, acrylamide hemoglobin adduct; HbGA, glycidamide hemoglobin adduct; MMA, methylmalonic acid; MUFA, sum of 6 monounsaturated fatty acids; PLP, pyridoxal-5'-phosphate; PUFA, sum of 11 polyunsaturated fatty acids; SFA, sum of 6 saturated fatty acids; sTfR, soluble transferrin receptor; tFA, total fatty acids [sum of 24 fatty acids]; tHcy, total homocysteine; uI, urine iodine; VIA, retinol; VIC, ascorbic acid; VIE, *alpha*-tocopherol; XAN, xanthophylls [sum of lutein, zeaxanthin and *beta*-cryptoxanthin]

[2] B, whole blood; P, plasma; S, serum; U, urine

Table 2

Mathematical forms of selected covariates for regression models

Chunk	Variable	Type	Transformation
Sociodemographic	Age, *y*	Continuous	None
	Sex	Categorical (2 levels)	N/A[1]
	Race-ethnicity	Categorical (5 levels)	N/A
	Poverty income ratio	Continuous	None
	Education	Categorical (2 levels)	N/A
Lifestyle	Smoking status	Categorical (2 levels)	N/A
	Alcohol consumption[2] *drinks/d*	Continuous	ln(x + 1)
	Supplement use	Categorical (2 levels)	N/A
	BMI (kg/m^2)	Continuous	ln
	Physical activity[3] *MET-min/wk*	Continuous	ln(x + 1)

[1] N/A, not applicable

[2] Alcohol consumption: calculated as average daily number of “standard” drinks [(quantity x frequency) / 365.25]; 1 drink ≈ 15 g ethanol

[3] Physical activity: calculated as total metabolic equivalent task (MET)-min/wk from self-reported leisure time physical activities

Table 3

Descriptive information for the adult US population ≥20 y by sociodemographic and lifestyle factors, NHANES 2003–2006

Factor	Category	Estimate[1]
Age, *y*	20–39	38.4
	40–59	38.8
	≥60	22.8
Sex	Male	48
	Female	52
Race-ethnicity	Mexican-American	7.9
	Non-Hispanic black	11.4
	Non-Hispanic white	72
	Other Hispanic	3.5
	Other (including multiracial)	5.4
Education	≤High school	44.2
	>High school	55.9
PIR[2]	Low	29.3
	Middle	28
	High	42.7
Smoking status[3]	No	71.2
	Yes	28.9
Alcohol consumption[4]	No drinks	29.4
	<1 (not 0)	56.8
	1–<2	7.9
	≥2	6.0
Supplement use[5]	No	45.9
	Yes	54.1
BMI[6]	Underweight	1.8
	Normal	31.6
	Overweight	33.4
	Obese	33.3
Physical activity[7]	None reported	32.1
	0–<500	24.2
	500–<1000	14.0
	≥1000	29.7

[1] Values represent weighted percentage using 4 y mobile examination center weights from NHANES 2003–2006

[2] PIR, family poverty income ratio; low: 0–1.85; medium: >1.85–3.5; high: >3.5

[3] "Smoker" defined by serum cotinine concentration >10 μg/L

[4] Alcohol consumption: calculated as average daily number of "standard" drinks [(quantity x frequency) / 365.25]; 1 drink ≈ 15 g ethanol

[5] "Supplement user" defined as participant who reported taking a dietary supplement within the past 30 d

[6] BMI (kg/m^2) definitions: underweight: <18.5; normal weight: 18.5–>25; overweight: 25–<30; and obese: ≥30

[7] Physical activity: calculated as total metabolic equivalent task (MET)-min/wk from self-reported leisure time physical activities

Table 4

Estimated percent change in biomarker concentration after adjusting for sociodemographic and lifestyle factors using data for adults ≥20 y, NHANES 2003–2006[1,2,3]

Analyte (matrix)[4]	Age: every 10 y increase	Sex: F vs. M[5]	Race-ethnicity: NHB vs. NHW[6]	Race-ethnicity: MA vs. NHW[6]	Education: ≤HS vs. >HS[7]	PIR[8]: every 2 unit decrease	Supplement use[9] yes vs. no	Smoking[10] yes vs. no	Alcohol[11] 1 vs. 0 drinks/d	BMI[12] 25% increase	Physical activity[13] 750 vs. 150 MET-min/wk
FOL (S)	6.8*	5.8*	−13.0*	−5.8*	−0.1	−1.3	38.4*	−14.9*	−2.4*	−4.1*	1.4*
FOL (RBC)	4.5*	3.6*	−19.7*	−4.7*	−0.7	−0.6	24.1*	−12.2*	1.6	3.8*	0.6*
PLP (S)	−2.1*	−21.2*	−7.7	8.1	−2.3	−8.3*	78.7*	−27.6*	10.6*	−12.6*	3.1*
4PA (S)	14.6*	−8.6*	−23.5*	−13.1*	−0.6	−6.3*	104*	−18.1*	−0.3*	−7.4*	1.3
B-12 (S)	1.0*	−3.7*	20.2*	15.4*	3.1	−0.9	20.8*	−6.2*	−3.3*	−4.3*	0.6
tHcy (P)	9.6*	−15.0*	0.3	−10.6*	−0.4	1.9*	−8.4*	7.8*	3.2*	−0.1	−0.4
MMA (P)	9.2*	−5.2*	−22.4*	−21.6*	−1.8	2.7*	−12.1*	−0.9	−1.9	−1.7	−1.4*
VIC (S)	1.9*	5.7*	3.4*	5.3*	−1.8*	−1.7*	16.2*	−11.0*	−0.8	−5.3*	1.4*
VIA (S)[14]	2.1*	−9.6*	−9.4*	−8.7*	−1.2	−0.7	5.4*	−0.3	6.1*	−1.4*	0.6*
VIE (S)[14]	5.0*	−1.2	−6.6*	5.2*	−2.8*	−2.3*	20.9*	−4.8*	−0.4	−0.9*	0.5
CAR (S)[14]	−2.8*	2.7	9.1*	−1.4	−9.3*	−3.7*	11.7*	−17.3*	0.1	−10.3*	2.8*
XAN (S)[14]	2.5*	−3.0*	33.1*	57.2*	−7.9*	−3.4*	6.1*	−24.8*	−1.6	−14.8*	2.5*
25OHD (S)[14]	−0.8*	0.1	−23.6*	−12.0*	1.6*	−0.6	5.2*	−1.5	1.6*	−4.3*	1.5*
SFA (P)[14]	0.4	2.0	−6.3*	8.4*	2.2	0.4	3.5*	1.5	4.8*	4.4*	−0.3
MUFA (P)[14]	2.4*	−1.2	−15.0*	10.0*	0.9	3.7*	2.0	5.4*	5.5*	4.2*	−0.5
PUFA (P)[14]	−0.1	1.8	0.4	8.9*	1.7	−1.6*	1.3	−2.4*	−0.6	0.3	0.3
tFA (P)[14]	0.7	0.6	−5.3*	10.1*	0.6	0.4	2.2	1.2	3.1*	2.0*	0.2
FER (S)	7.6*	N/A[15]	−6.6	−3.9	−5.0	−4.2	6.0	24.2*	21.3*	7.2*	2.9*
sTfR (S)	1.5	N/A	18.8*	−3.9	−2.6	3.4*	−1.1	−10.7*	−5.9*	6.0*	−0.7
BI (S)	0.2*	N/A	−0.8*	0.0	−0.1	−0.3*	0.3	1.1*	0.9*	0.0	0.1
uI (U)[16]	11.4*	3.8	−33.7*	3.7	4.6	3.4	22.1*	−6.6	−7.1*	−0.5	0.1

Analyte (matrix)[4]	Age: every 10 y increase	Sex: F vs. M[5]	Race-ethnicity: NHB vs. NHW[6]	Race-ethnicity: MA vs. NHW[6]	Education: ≤HS vs. >HS[7]	PIR[8]: every 2 unit decrease	Supplement use[9] yes vs. no	Smoking[10] yes vs. no	Alcohol[11] 1 vs. 0 drinks/d	BMI[12] 25% increase	Physical activity[13] 750 vs. 150 MET-min/wk
EQU (U)[16]	0.9	19.5*	−41.2*	−35.7*	1.1	−7.9	13.6	−6.5	−17.8*	−2.5	4.3*
DMA (U)[16]	7.7*	28.0*	−17.5	−55.6*	−9.3	−14.6	12.0	−28.3*	−20.2*	−3.1	2.3
ETD (U)[16]	8.7*	39.7*	−24.1*	8.9	−15.6	−16.5*	8.6	−10.5	12.9	−10.8*	2.9
ETL (U)[16]	12.4*	15.5	−17.6	24.4	−6.5	−13.7*	1.0	−32.3*	−5.4	−21.1*	6.5*
DAZ (U)[16]	9.2*	10.4	−15.0*	−16.6*	−9.5	−3.8	3.3	−12.8	−6.6	−3.4	0.9
GNS (U)[16]	9.5*	13.6	−22.8*	−4.4	−7.6	−2.1	10.3	−8.9	−2.4	−8.1	1.9
HbAA (B)	−2.9*	3.6	0.1	4.0	5.9	−0.1	−3.5*	126*	2.4	−4.8*	0.6
HbGA (B)	−5.7*	8.7*	16.9*	4.7	7.0	−2.0	−2.5	101*	−11.8*	2.6*	0.2

[1] Change represents percent change (%) in geometric mean for all biomarkers except for vitamin C (μmol/L), 25-hydroxy vitamin D (μmol/L) and body iron (mg/kg) where change in arithmetic mean represents concentration units; change in each covariable was carried out while holding any other variables in the model constant; 25OHD, 25-hydroxyvitamin D; 4PA, 4-pyridoxic acid; B-12, total cobalamin; BI, body iron; CAR, carotenes [sum of *alpha*-carotene, *beta*-carotene and *cis*- and *trans*-lycopene], DAZ, daidzein ; DMA, O-desmethylangolensin; EQU, equol; ETD, enterodiol; ETL, enterolactone; FER, ferritin; FOL, folate; GEN, genistein; HbAA, acrylamide hemoglobin adduct; HbGA, glycidamide hemoglobin adduct; MMA, methylmalonic acid; MUFA, sum of 6 monounsaturated fatty acids; PLP, pyridoxal-5'-phosphate; PUFA, sum of 11 polyunsaturated fatty acids; SFA, sum of 6 saturated fatty acids; sTfR, soluble transferrin receptor; tFA, total fatty acids [sum of 24 fatty acids]; tHcy, total homocysteine; uI, urine iodine; VIA, retinol; VIC, ascorbic acid; VIE, *alpha*-tocopherol; XAN, xanthophylls [sum of lutein, zeaxanthin and *beta*-cryptoxanthin]

[2] Iron status indicators (FER, sTfR and BI) were only measured in women of reproductive age, thus our analysis was limited to women 20–49 y of age

[3] Hb AA, HbGA, MMA, SFA, MUFA, PUFA, and tFA, data only available for NHANES 2003–2004; 4PA, CAR, PLP, VIA, VIE, and XAN data only available for NHANES 2005–2006

[4] B, whole blood; P, plasma; S, serum; U, urine

[5] F, female; M, male

[6] MA, Mexican American; NHB, non-Hispanic black; NHW, non-Hispanic white

[7] HS, high school

[8] PIR, family poverty income ratio

[9] "Supplement user" defined as participant who reported taking a dietary supplement within the past 30 d

[10] "Smoker" defined by serum cotinine concentration >10 μg/L

[11] Alcohol consumption: calculated as average daily number of "standard" drinks [(quantity x frequency) / 365.25]; 1 drink ≈ 15 g ethanol

[12] A 25% increase in BMI is comparable to a change from being normal weight to overweight

[13] Physical activity: calculated as total metabolic equivalent task (MET)-min/wk from self-reported leisure time physical activities

[14] Model includes total cholesterol and lipid altering prescription drug use

[15] N/A; not applicable because data were only available for women

[16] Model includes urine creatinine

* Change is significantly different from 0; $P<0.05$